15-440/640 Distributed Systems
Midterm II SOLUTION

Name:
Andrew ID:

December 15, 2016

- Please write your name and Andrew ID above before starting this exam.
- This exam has 11 pages, including this title page. Please confirm that all pages are present.
- This exam has a total of 100 points.

Question	Points	Score
1	12	
2	22	
3	14	
4	22	
5	14	
6	15	
7	1	
Total:	100	

True/False

1. (12 points) Indicate whether each statement below is *true* or *false*. ***ALSO****, give a one sentence reason for your answer.

   Each correct answer is worth 3 points

   A. [True, False] For a particular Map-Reduce Phase with 'M' and 'R' nodes respectively, each reducer node handles 1/R of the intermediate key-values by default.

   Solution: TRUE. Each Reducer node handles 1/R of the possible key-space and fetches them from each of the M Mappers.

   B. [True, False] In case of System Virtualization, the hypervisor (or VMM) can only allocate and assign the actual number of physical processor resources it has available at its disposal to individual Virtual Machines (VMs).

   Solution: FALSE. Virtualization allows you to overcommit resources and provide more virtual resources than actual physical resources, for example by time multiplexing the same set of CPU/processors across multiple VMs .

   C. [True, False] DES (symmetric key) is faster than RSA (public-private key)

   Solution: Ans: **TRUE**. Symmetric key ciphers are generally going to be faster than Public-private key ciphers due to the mathematical complexity of the latter.

   D. [True, False] When setting up a TOR circuit the sender is guaranteed that the path taken (intermediate routers) to the receiver will be the same as without using TOR.

   Solution: Ans: **FALSE**. There are no guarantees, and the sender decides which path to take when using TOR, which comprises of TOR nodes and not the regular internet routers.

Short Answers

2. In the following, keep your answers brief and to the point (i.e. 1-2 sentences).

   (a) (4 points) Consistent hashing is used in a variety of distributed systems over more traditional hashing algorithms.

      A. Describe one way that consistent hashing performs better when failures occur (e.g. in DHTs or in CDNs).

      Solution: Less content needs to be moved between nodes.

      B. Describe one way that consistent hashes perform better when nodes have differing views on the availability of servers.

      Solution: Consistent hashing reduces the spread of requests across nodes when views are different

   (b) (4 points) You are working on a MAC laptop and using it to develop and test a new Linux OS/kernel. You are worried about bugs since they will cause the Linux kernel to crash. Explain how the 'Isolation' property of hosted Virtualization solution like VMWare Workstation running on your MAC can help in this case.

   Solution: Fault Isolation: We can run the new Linux kernel in a VM. If this VM kernel contains a bug, and crashes it will not crash the main host (OSX) thereby providing fault isolation and allow development and testing to continue.

(c) (6 points) We studied the MapReduce programming paradigm for large scale data analysis on clusters, and compared and contrasted that to the more traditional Bulk Synchronous Programming (BSP) methods.

A. Explain what the problem with 'stragglers' is when using MapReduce and how is it addressed?

> **Solution:** Stragglers mean that some Map/Reduce computations may take much longer than others, delaying the entire phase. Solution is to rerun the tasks on nodes that have finished first and then take the result for that task from whoever finishes first.

B. Under the MapReduce programming model, how does one deal with a single master node failure? (Assume there is a single master node).

> **Solution:** Master failure (discussed in class and also in the paper assigned for readin). 1. Master writes periodic checkpoints such that a new copy can be started from the last checkpointed state when master fails. 2. If only a single master, that particular MapReduce task is aborted and has to be restarted with a new master.

(d) (4 points) Give one attack that would be possible if self-certifying names did not use cryptographic hashes.

> **Solution:** If you used a generic hash, you could use generate an alternate file with the same hash value. This would then be accepted as legitimate content.

(e) (4 points) The DNS root and GTLD servers often come under attack. Give 2 reasons that we don't notice when they are under attack.

> **Solution:** The records from the root and GTLD have long TTLs and are cached widely.
>
> The root and GTLD zones are replicated very widely and it is difficult to attack all locations at the same time.

Byzantine Generals vs. Pirates

3. Srini decides to outsource his distributed system nodes to NeverNeverLand. Help him understand the reliability/availability tradeoffs of his system design.

   (a) (3 points) The LostBoys Corporation runs Srini's 61 server nodes. While the LostBoys run a safe system (i.e. nodes are never corrupted or run faulty code), they may fail (i.e. fail-stop) occasionally. Srini decides to use two-phase commit in his distributed application. If Srini wants to provide 100% availability to his customers, what guarantee does he need from LostBoys (i.e. how many LostBoy server nodes can fail)? Give a 1-2 sentence explanation.

   Solution: all nodes must be reliable – i.e. no failures. If there is one failuer, the two phase commit will fail

   (b) (3 points) The LostBoys Corporation runs Srini's 61 server nodes. Srini decides to upgrade to using Paxos in his distributed application. If Srini wants to provide 100% availability to his customers, what guarantee does he need from LostBoys (i.e. how many LostBoy server nodes can fail)? Give a 1-2 sentence explanation.

   Solution: upto 30 nodes can fail - paxos uses majority voting

   (c) (4 points) It turns out that the LostBoys occasionally outsource the operation of some of the server nodes to HookCo. Unfortunately, HookCo is notoriously untrustworthy and their servers suffer from all types of failures and are often running corrupted code. The LostBoys promise that despite the use of HookCo nodes only a limited fraction of Srini's 61 total nodes will fail at any time (no more than your answer to part b). If Srini continues to run Paxos on his nodes, will his system work fine? Why or why not?

   Solution: Paxos may produce inconsistent results since it is not designed to deal with nodes that don't follow the protocol correctly

   (d) (4 points) On the set of 61 nodes (some run by LostBoys and some run by HookCo), Srini decides to use the PBFT (Practical Byzantine Fault Tolerance) protocol in his distributed application. If Srini wants to provide 100% availability and correct results to his customers, what guarantee does he need from the set of 61 total server nodes? Give a 1-2 sentence explanation.

   Solution: PBFT requires 3f+1 nodes – therefore f = 20 – no more than 20 nodes can fail/not follow protocol at any time. This probably means that < 20 HookCo nodes and at most 20-#hookco nodes can failstop at any time

DHTs, CDNs & DNS - our favorite TLAs (three letter acronyms)

4. As the CMU's Distributed Systems class graduates ventured into the world, they spread their description of class lectures. As a result, the popularity of video recordings of the lectures skyrocketed (unfortunately, mostly for the jokes rather than the technical content) and now Srini needs your help in managing the storage and delivery of the files.

   Srini starts by building an Akamai-like CDN to deliver the files for his web site www.srini-lectures.com. As a first step, Srini:

   1. registers the domain srini-lectures.com with two name servers ns1.srini-lectures.com and ns2.srini-lectures.com.
   2. sets up two regional, low-level DNS servers - east-ns.srini-lectures.com and west-ns.srini-lectures.com - for clients on the east and west coast.
   3. sets up two servers to deliver cached content with IP addresses 1.1.1.1 and 2.2.2.2. Srini expects users to get the 1.1.1.1 address for www.srini-lectures.com if they are on the east coast and 2.2.2.2 if they are on the west coast.

   (a) (9 points) There are 6 name servers involved in the system:
   a.root-servers.net
   a.gtld-servers.net
   ns1.srini-lectures.com
   ns2.srini-lectures.com
   east-ns.srini-lectures.com
   west-ns.srini-lectures.com

   For each of the following DNS records: 1) list *ALL* servers where the DNS record stored (do not list cached locations), 2) what will its TTL be (assume that the TTL can only be 1 minute or 1 day) and 3) why you chose that TTL (1-2 sentences)

   A. NS record that points to ns1.srini-lectures.com

   Solution: on a.gtld-servers.net with TTL 1 day

   B. NS record that points to east-ns.srini-lectures.com

   Solution: on ns1.srini-lectures.com and ns2.srini-lectures.com with TTL 1 day

   C. A record for www.srini-lectures.com

   Solution: on east-ns.srini-lectures.com and west-ns.srini-lectures.com with TTL 1 min

(b) (3 points) On the first query (i.e. no cached information at the local name server) for www.srini-lectures.com from New York - which of the 6 name servers listed are contacted?

Solution: a.root-servers.net
a.gtld-servers.net
ns1.srini-lectures.com or ns2.srini-lectures.com east-ns.srini-lectures.com

(c) (3 points) On a second query, 5 minutes after the previous query, for www.srini-lectures.com from New York - which of the 6 name servers listed are contacted?

Solution: east-ns.srini-lectures.com

Srini decides to move the content to a Chord DHT. He splits the lectures in 10 second chunks, which results in 65536 (2^{16}) files. He uses SHA3-256 (256bit long hash) to generate the identifiers used in his DHT. Finally, he recruits 1024 (2^{10}) users to participate in the Chord DHT as storage nodes.

(d) (4 points) How many hops does it typically take to lookup a single chunk?

Solution: log(1024) = 10 hops

(e) (3 points) How many different hosts are typically in the Chord finger table at any node?

Solution: log(1024) = 10 nodes

That's Where Yahoo Came From?

5. Help Gulliver understand his storage system properties. During his travels, Gulliver hears that the tiny workers at LiliputianCo have great attention to detail and maintain their systems well. They provide drives with the following specifications:

MTTF	72 years	
Sequential read/write speed	100 MB/sec	(100×10^6 bytes/sec)
Capacity	1 TB	(1×10^{12} bytes)
I/O per second	100,000	

(a) (3 points) Because the Liliputian drives are so reliable, Gulliver decides to use a simple configuration for storage. He starts by striping his data across 8 drives. In this configuration, calculate how long before Gulliver is likely to lose data (i.e. MTDL—mean time to data loss).

Solution: 72/8 = 9 years

(b) (3 points) Gulliver's gets nervous about reliability and decides to add a parity disk to his system. The end configuration uses RAID-4 (striping with a parity disk) with 8 data drives and 1 parity disk. In this configuration, calculate how long before Gulliver is likely to lose data (i.e. MTDL—mean time to data loss).

Solution: 72/9 + 72/8 = 9+8 = 17 years for a set

(c) (4 points) Even though there are no failures, Gulliver feels that adding the parity disk slowed down his performance. What types of operations are probably slower on his new array and why?

Solution: small writes are probably the worst – they will be twice the latency or use 4 times the BW since they require a read-modify-write operation.

(d) (4 points) Gulliver is still not satisfied with the reliability. He hires YahooCorp to monitor his drive array and replace a drive. Unfortunately, Yahoos are notoriously unreliable and it takes them 6 months to repair a broken drive on average. In this configuration, calculate how long before Gulliver is likely to lose data (i.e. MTDL—mean time to data loss).

Solution: 72/9 * (72/8)/6month = 9*8yr/6month = 9 * 16 = 144 years

Crypto Scenarios – because you know, the Internet is a bad place!

6. (15 points) Below are several scenarios describing simple uses of cryptographic schemes we have covered in class. For each scenario, circle "correct" if the scenario describes a valid use of the mechanism as described in class. Otherwise circle "incorrect". In either case, provide *one sentence* explaining the vulnerability it exposes or why there is no vulnerability.

   (a) (3 points) Yuvraj wants to transmit P3 grades from his home computer to Srini at CMU. He is worried that an enterprising 440/640 student may have hacked a router along the path and might modify the message to improve their grade and win the P3 App contest. So when Yuvraj sends a message M to Srini, he also calculates H = Hash(M) and appends H to the message. Srini receives both M and H, and calculates H' = Hash(M), only accepting the message as valid if H' = H .(Assume that Hash is a secure hash function with the standard cryptographic hash properties. Also, Srini and Yuvraj do not have any shared secret keys.)

   correct / incorrect

   Solution: Incorrect. The enterprising student could see and change the contents and recalculate the hash since we don't use encryption for confidentiality. Srini would never know that anything had changed.

   (b) (3 Points) Srini now wants to send the top-secret solutions for the final exam to Yuvraj, but Yuvraj's email server is down. However, Yuvraj checks the public 440 Piazza site regularly to see if Srini posted any messages for him. During the first day of class, Yuvraj personally gave everyone at the lecture (including Srini) his public key K_{Yuvraj}. Srini, similarly gave everyone his public key K_{Srini}. Only Yuvraj or Srini know their private keys, K^{-1}_{Yuvraj} and K^{-1}_{Srini} respectively. Srini then calculates a signature S for the answer key using K^{-1}_{Srini}, and then encrypts both this signature S and the answer key with K_{Yuvraj}, and then posts it to Piazza.

   correct / incorrect

   Solution: Correct. Since Srini used asymmetric keys, specifcially the public key for Yuvraj only Yuvraj can decrypt the message. Also, since the answer key is signed using Srini's private key, Yuvraj can verify using his public key that the solution came from Srini.

(c) (3 points) Yuvraj and Srini learned about KDCs and want to give that a try, and both of them share a secret key with a Key Distribution Center (KDC). We call these keys $K_{Yuvraj,KDC}$ and $K_{Srini,KDC}$ respectively. Yuvraj wants to establish a shared symmetric key with Srini, so yuvraj authenticates to the KDC using $K_{Yuvraj,KDC}$ and the KDC replies with $Encrypt_{K_{Yuvraj,KDC}}(K_{Srini,KDC})$. Yuvraj and Srini then communicate using the shared secret key $K_{Srini,KDC}$.

correct / incorrect

Solution: Incorrect. Yuvraj would now be able to pretend to be Srini because he has the secret key that Srini shares with the KDC. Instead they should have used a session key/ticket.

(d) (6 points) Srini and Yuvraj realize that also need a mechanism to communicate and decide which TA is going to grade the next homework. They have a shared secret, K_{Profs} that allows them to create unforgable message authentication codes (MAC) so that Yuvraj can verify that Srini did in fact create any message that is received. Yuvraj and Srini have a simple protocol: Yuvraj sends a "Who grades HWX?" message to Srini in plain text, and Srini replies with one of three messages: $M1 = MAC_{K_{Profs}}$("TA-1"), $M2 = MAC_{K_{Profs}}$("TA-2"), or $M3 = MAC_{K_{Profs}}$("TA-3"). When Yuvraj receives either M1, M2, or M3, he verifies the MAC using K_{Profs} and knows who will grade the next homework.

(Part A) This protocol is insecure. A malicious TA could avoid grading anything! Explain the attack?

Solution: (Part A) It is subject to a replay attack. The TA could replay an earlier answer for a different TA's name.

(Part B) What simple change to the above protocol could defend against this attack?

Solution: (Part B) Use a nonce. Along with the "Who grades HWX" message, Bob should also send a random string that must be included in the MAC to ensure that the answer is unique.

Anonymous Feedback

7. (1 point) Tear this sheet off to **receive one point**. We'd love it if you handed it in either at the end of the exam or, if time is lacking, to the course secretary.

   (a) Please list one thing you'd like to see improved in this class in the current or a future version.

   (b) Please list one good thing you'd like to make sure continues in the current or future versions of the class.